Turtle Games Report

Turtle Games is a game manufacturer and retailer boasting a global customer base. The company not only manufactures and sells its own products but also sources and sells products sold by other companies. Products in the company's catalogue include books, board games, video games, and toys. Turtle Games has the objective of improving overall sales performance by utilising customer trends. This report will focus on analysing different aspects of the company and helping Turtle Games achieve this objective.

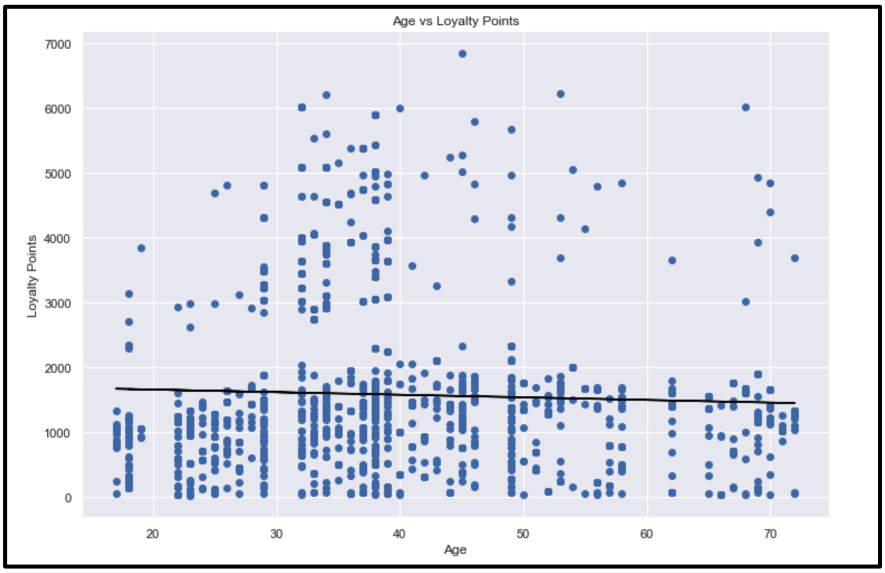
I began by first exploring the dataset 'turtle_reviews.csv' to obtain a general understanding. These included checking the data types of the different columns and inspecting for any missing or duplicated values. I also cleaned the dataset a little by changing column names for easier reference.

I started by analysing how the customers of Turtle Games accumulated loyalty points. I ran linear regressions between the dependent variable, loyalty_points and the independent variables age, renumeration and spending_score to check for any relationships between these variables.

Dep. Variable:	y	**R-squared:**	0.002
Model:	OLS	**Adj. R-squared:**	0.001
Method:	Least Squares	**F-statistic:**	3.606
Date:	Tue, 20 Dec 2022	**Prob (F-statistic):**	0.0577
Time:	13:40:31	**Log-Likelihood:**	-17150.
No. Observations:	2000	**AIC:**	3.430e+04
Df Residuals:	1998	**BIC:**	3.431e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	1736.5177	88.249	19.678	0.000	1563.449	1909.587
X	-4.0128	2.113	-1.899	0.058	-8.157	0.131

Omnibus:	481.477	**Durbin-Watson:**	2.277
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	937.734
Skew:	1.449	**Prob(JB):**	2.36e-204
Kurtosis:	4.688	**Cond. No.**	129.

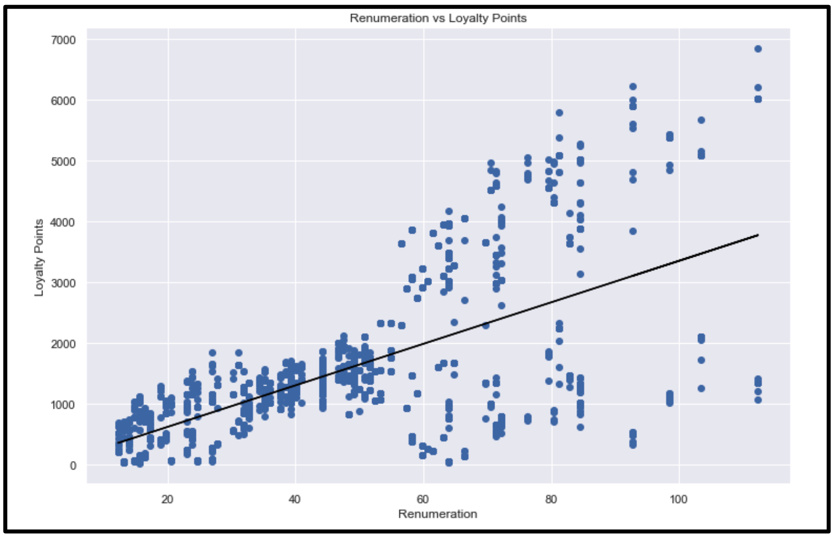


The figures above show the OLS regression results for the regression run on Age vs Loyalty Points. From looking at the scatterplot, we can see that there is extremely weak negative correlation which is further evidenced by the weak R-squared value of 0.002. Furthermore, we can see that the p value of 5.8% is greater than 5%, deeming the relationship between the variables statistically insignificant. The regression line also has a poor goodness of fit meaning we can consider the variable 'age' as an unsuitable variable for predicting loyalty points.

Dep. Variable:	y1	R-squared:	0.380
Model:	OLS	Adj. R-squared:	0.379
Method:	Least Squares	F-statistic:	1222.
Date:	Tue, 20 Dec 2022	Prob (F-statistic):	2.43e-209
Time:	13:40:32	Log-Likelihood:	-16674.
No. Observations:	2000	AIC:	3.335e+04
Df Residuals:	1998	BIC:	3.336e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-65.6865	52.171	-1.259	0.208	-168.001	36.628
X1	34.1878	0.978	34.960	0.000	32.270	36.106

Omnibus:	21.285	Durbin-Watson:	3.622
Prob(Omnibus):	0.000	Jarque-Bera (JB):	31.715
Skew:	0.089	Prob(JB):	1.30e-07
Kurtosis:	3.590	Cond. No.	123.

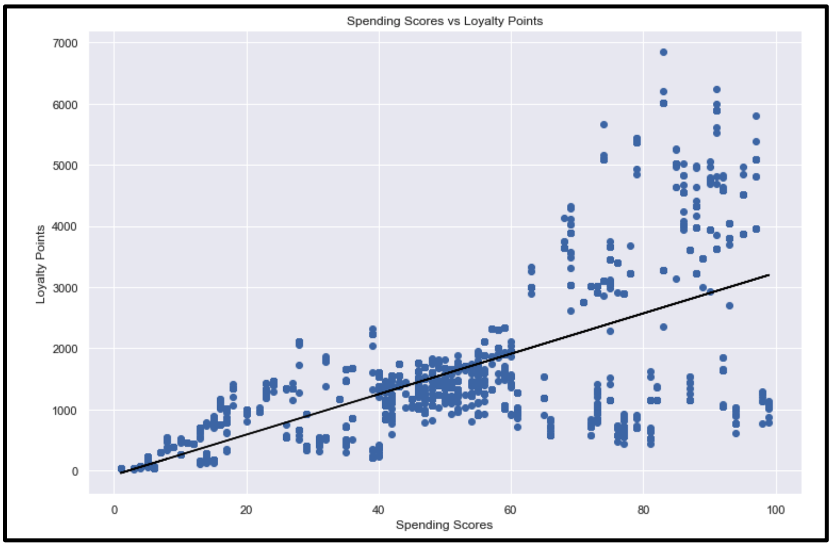


The figures above show the relationship between the Renumeration and Loyalty points variables. From the scatterplot, we can see strong positive correlation which is evidenced by the R-squared value being 0.380. The p value is less than the 5% significance level which indicates the relationship between the variables is statistically significant. The same can be said for the relationship between the 'spending score' and 'loyalty points' variable, seen in the graphs below, which has a R-squared value 0.452 and is also statistically significant.

Dep. Variable:	y2	R-squared:	0.452
Model:	OLS	Adj. R-squared:	0.452
Method:	Least Squares	F-statistic:	1648.
Date:	Tue, 20 Dec 2022	Prob (F-statistic):	2.92e-263
Time:	13:40:33	Log-Likelihood:	-16550.
No. Observations:	2000	AIC:	3.310e+04
Df Residuals:	1998	BIC:	3.312e+04
Df Model:	1		
Covariance Type:	nonrobust		

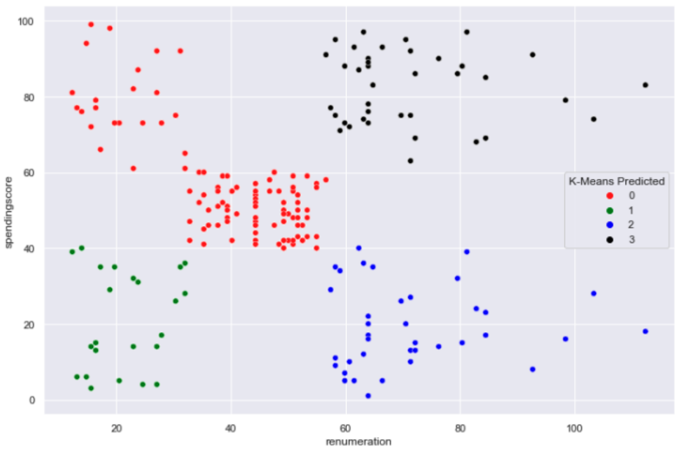
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-75.0527	45.931	-1.634	0.102	-165.129	15.024
X2	33.0617	0.814	40.595	0.000	31.464	34.659

Omnibus:	126.554	Durbin-Watson:	1.191
Prob(Omnibus):	0.000	Jarque-Bera (JB):	260.528
Skew:	0.422	Prob(JB):	2.67e-57
Kurtosis:	4.554	Cond. No.	122.

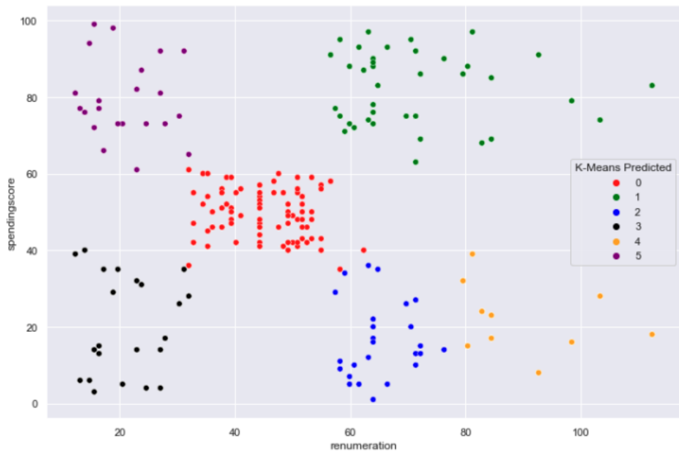


I next studied how groups within the customer base can be used to target specific market segments. I utilised k-clustering to identify the optimal number of clusters and then applied and plotted the data using the created segments. I utilised the elbow method and the silhouette method to determine the optimal cluster value and decided to investigate 3 possible values. The three values for k are seen below in 3 different scatterplots using seaborn.

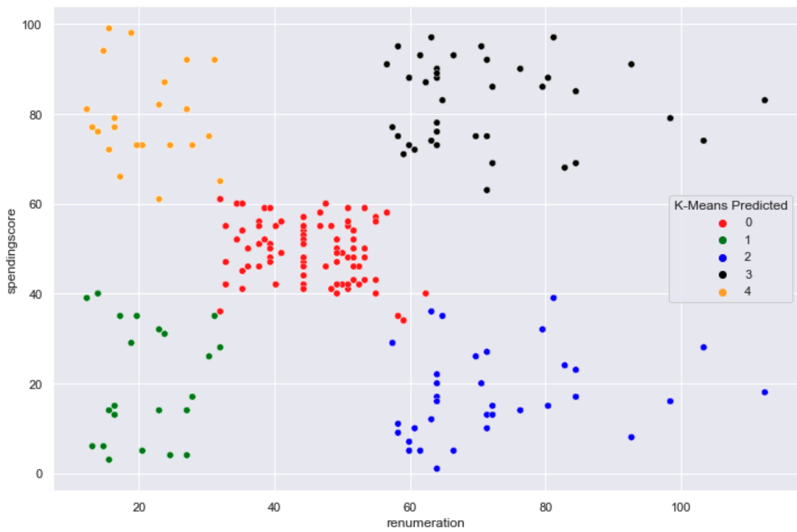
K= 4



K = 6



K=5

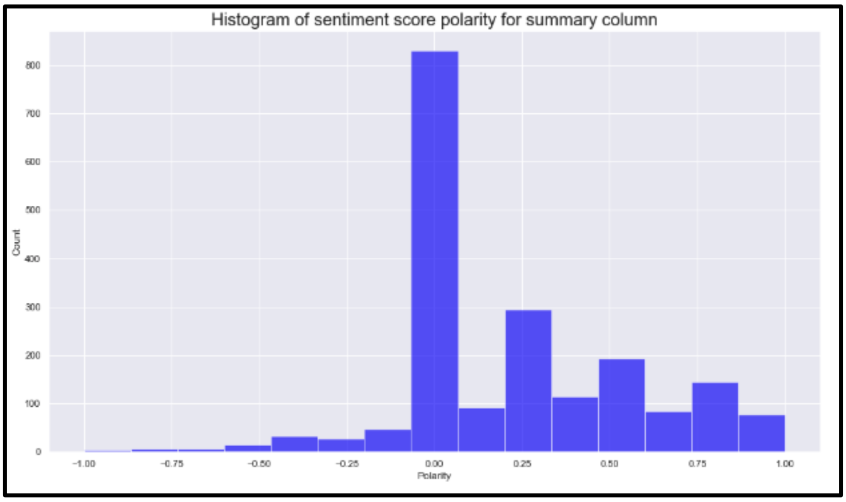


Looking at the K=4 graph, we can see it isn't optimal due to overlap between different clusters caused by there being too few clusters. With the K = 6 graph, we can see that this also isn't optimal as there are too many clusters that would overcomplicate the analysis. Only the K=5 graph shows the optimal number of clusters.

Analysing the 5-cluster scatterplot, we can suggest the Turtle Game marketing team to focus on customers within the green and blue clusters. This is because despite them having similar renumeration to the yellow and black clusters, they have much lower spending scores compared to these two clusters. The marketing team would have to study why these two clusters spend so low and use that information to boost their spending.

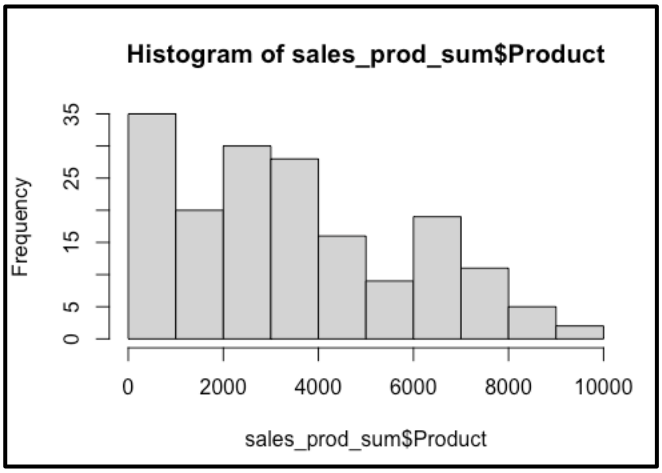
I next studied how social data could be expended to inform marketing campaigns by analysing customer sentiments with reviews. I conducted my analysis by creating a separate dataframe to focus on the ‘review’ and ‘summary’ columns, which I then tokenized as part of Natural Language Processing to identify the top 15 most common words followed by the top 20 positive and negative reviews and summaries based on their respective sentiment polarities.

Above is the word cloud I created using the most commonly occurring words in the ‘summary’ column. Using buzzwords such as ‘five stars’, ‘excellent’ and ‘fun’ could help Turtle Games further increase their sales as these words can be used in tangent with popular items for customers such as ‘game’ and ‘book’.

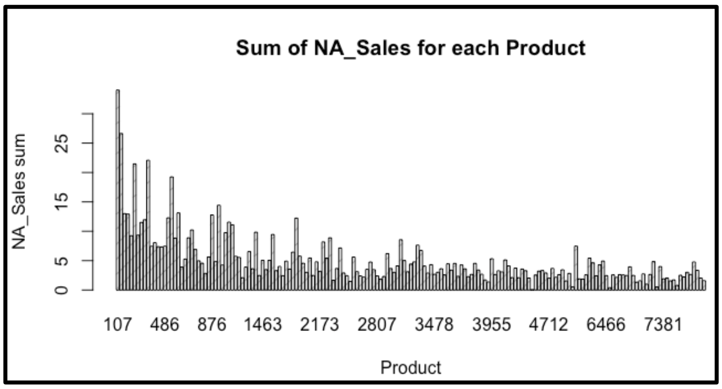


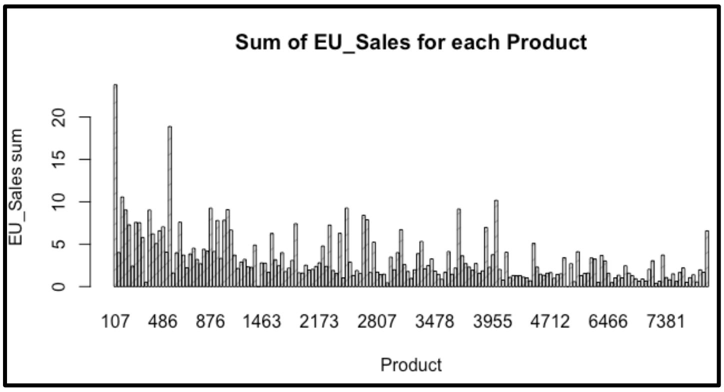
The histogram shows that the sentiment score polarity for the summary column is weakly skewed to the right where the mean > median > mode. The most frequent summary is a neutral one which could be due to misinterpretations of certain words used in summaries. The review histogram is however negatively skewed with the most popular review being a positive one. This can give the company an indication of the things going well and what should be maintained compared to the negative reviews which can show what needs to be improved on.

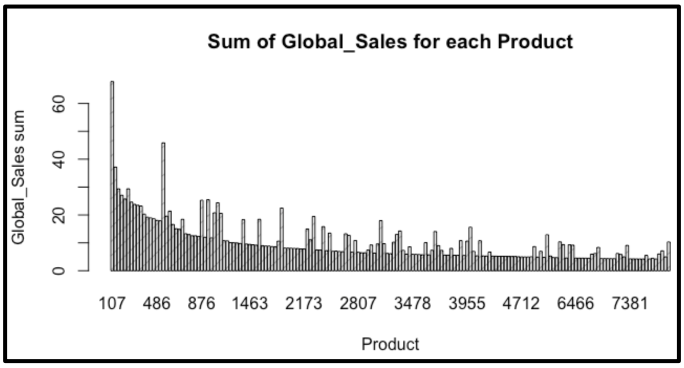
The next step of my analysis was to look at the impact that each product has on sales using R.



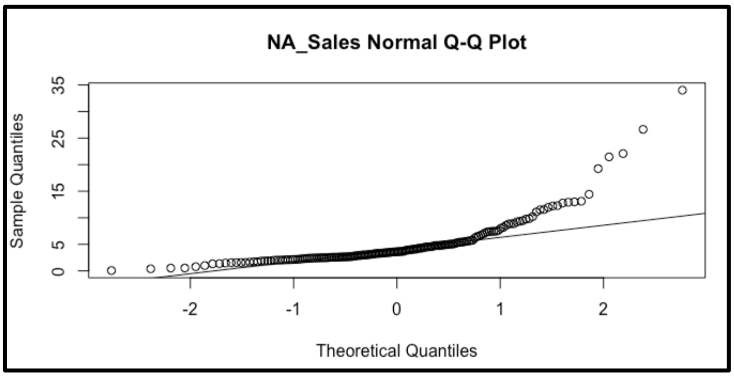
The histogram above shows the frequency sold of each product and is positively skewed. Products with lower product IDs are sold in higher frequencies. Turtle Games could use this information to try find reasons why products with higher IDs sell less.

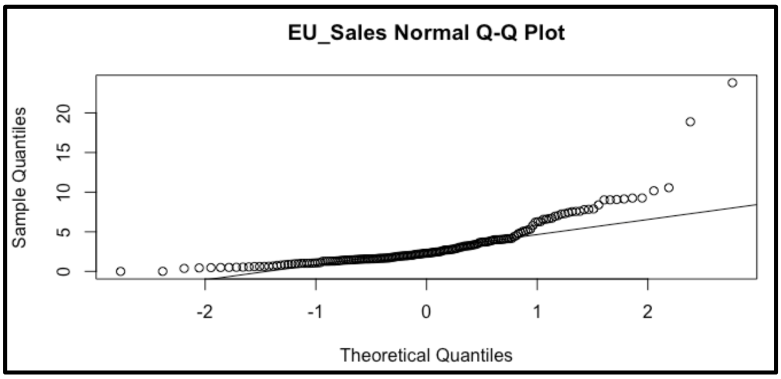


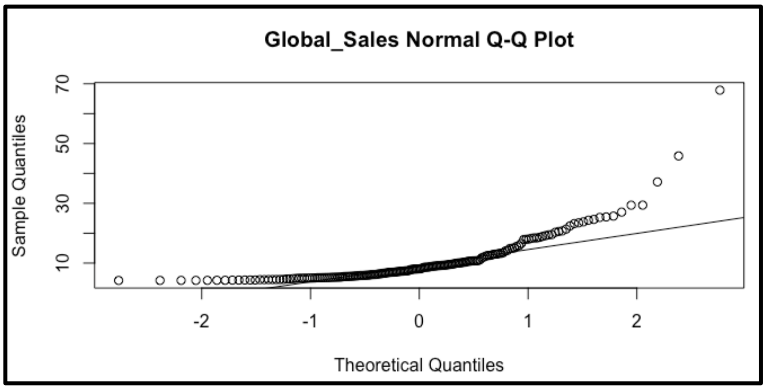




The graphs above confirm the trend shown by the histogram where products with lower product IDs have higher sales compared to those with higher product IDs.







The QQ plots shown above all show positively skewed data as the upper end of the Q-Q plots deviate from the straight line and the lower end follows a straight line for each of the three plots.

The trend in skewness of the data is further demonstrated through the results occurring from the Skewness and Kurtosis tests shown below.

```
#Important to note that:
#skewness > 1 indicates highly skewed data and
#kurtosis < 2 indicates normal distribution.

# Skewness and Kurtosis for NA_Sales.
skewness(sales_prod_sum$NA_Sales) # = 3.048198>1, so highly skewed data.
kurtosis(sales_prod_sum$NA_Sales) # = 15.6026>2, so not normally distributed.

# Skewness and Kurtosis for EU_Sales.
skewness(sales_prod_sum$EU_Sales) # =  2.886029>1, highly skewed data.
kurtosis(sales_prod_sum$EU_Sales) # = 16.22554>2, not normally distributed.

# Skewness and Kurtosis for Global_Sales.
skewness(sales_prod_sum$Global_Sales) # = 3.066769>1, highly skewed data.
kurtosis(sales_prod_sum$Global_Sales) # = 17.79072>2, not normally distributed.
```

I decided to conclude my analysis by determining if there were any relationships between North American, European and Global Sales. I ran a model that studied the relationships through multiple linear regression and discovered positive correlation between all three variables. This is seen in the figure below.

```
# a) NA_Sales_sum = 34.02 and EU_Sales_sum = 23.80
Global_predicted <- data.frame(NA_Sales=c(34.02), EU_Sales=c(23.8))
predict(GL_EU_NA, newdata=Global_predicted)
# Observed value =67.85, predicted value = 68.06

# b) NA_Sales_sum = 3.93 and EU_Sales_sum = 1.56
Global_predicted <- data.frame(NA_Sales=c(3.93), EU_Sales=c(1.56))
predict(GL_EU_NA, newdata=Global_predicted)
# Observed = 6.04, predicted = 7.36

# c) NA_Sales_sum of 2.73 and EU_Sales_sum of 0.65
Global_predicted <- data.frame(NA_Sales=c(2.73), EU_Sales=c(0.65))
predict(GL_EU_NA, newdata=Global_predicted)
# Observed = 4.32, predicted = 4.908

# d) NA_Sales_sum of 2.26 and EU_Sales_sum of 0.97
Global_predicted <- data.frame(NA_Sales=c(2.26), EU_Sales=c(0.97))
predict(GL_EU_NA, newdata=Global_predicted)
# Observed = 3.53, predicted = 4.76

# e) NA_Sales_sum of 22.08 and EU_Sales_sum of 0.52
Global_predicted <- data.frame(NA_Sales=c(22.08), EU_Sales=c(0.52))
predict(GL_EU_NA, newdata=Global_predicted)
# Observed = 23.21, predicted = 26.626
```

We can conclude that there is positive correlation with the variables loyalty_points, renumeration and spending_score. An increase in either of the latter two factors leads to an increase in loyalty_points.

Studying the clusters, we identified the groups of the customer base where spending scores were low for the company to show extra focus towards. Ideal cluster was when K=5.

Using the sentiment analysis, the company can deduce where things are going well by the positive reviews and see where things aren't going through negative reviews. Certain buzzwords seen in the world cloud can be used more frequently to attract customers.

Moving onto product analysis, we discovered that those with lower IDs sold more than those with higher product IDs, which the company could dwell further into. We also found a positive correlation between different sales regions. Global_Sales would increase whenever NA_Sales and EU_Sales increased.

Finally, we saw that the data sets were all skewed meaning there was a lack of normal distribution. This can easily be combatted by converted using logarithmic transformation which would convert the dataset to a normal distribution.